Additional file 3

The adaptive community-response (ACR) method for collecting misinformation on social media

Julian Kauk[1], Helene Kreysa[1], André Scherag[2, 3], and Stefan R. Schweinberger[1, 3, 4]

[1] *Department of General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Jena, Germany*
[2] *Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Jena, Germany*
[3] *Michael Stifel Center Jena, Jena, Germany*
[4] *German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany*

August 29, 2023

Additional indicators validating the ACR method

Here, we present a set of additional indicators validating the ACR method. These indicators are mainly based on a correlative approach and can be considered as logical consequences of the mathematical framework proposed in the main text.

Association between recall and precision

Initially, we tested the hypothesis whether there is a (negative) association between recall and precision, as implicated by the precision-recall tradeoff. We found that recall and overall precision were *not* negatively correlated ($r(346) = .063, p = .878, \mathrm{one-tailed}$; see Fig. 1a), contradicting our hypothesis. The absence of this correlation, however, can be explained by the low variance of precision: As we (intentionally) bounded the overall precision to $[0.95, 1]$, a potential correlation may have been suppressed. However, the distribution of recall (see Fig. 2c in the main text) may be considered as a signature of the precision-recall tradeoff. As stated previously, we observed a low median recall, arguably due to the high precision threshold. Furthermore, the positive skewness of the distribution may also support the assumption that both precise and sensitive queries can *not* occur frequently.

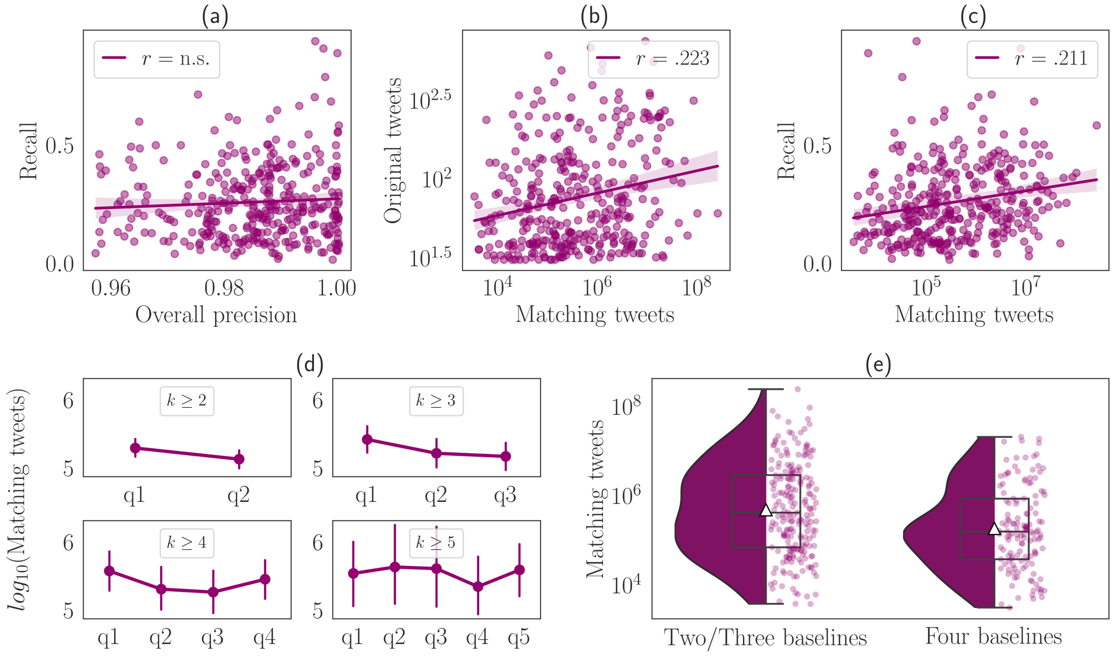

Figure 1: Additional indicators of the ACR method. Please note that (i) overall precision and recall are estimates, (ii) the number of matching tweets was log_{10}-transformed for all analyses, and (iii) regression lines reflect ordinary least squares regression, with 95% confidence interval (estimation via bootstrapping). (a) No significant negative correlation was observed between recall and overall precision. (b) The number of fact-checking tweets and the log_{10}-transformed number of fact-checked tweets were positively correlated, indicating that larger stories also mobilize greater fact-checking efforts. (c) The number of matching tweets showed a positive association with recall, indicating that more sensitive queries also tended to match more tweets. (d) Associations between subquery position and recall. Each panel reflects the average number of matching tweets, given a minimum number of subqueries and different subquery positions. Error bars reflect 95% confidence interval. There is an evident decline in the number of matching tweets for stories having at least two or three subqueries (top panels), whereas such effects were not so pronounced or not present for a higher number of minimal subqueries (bottom panels). (e) Number of matching tweets given different numbers of baselines. The number of matching tweets was reduced for four baselines, indicating that more baselines lead to an overload reduction.

Association between the number of matching tweets and the number of fact-checked tweets

We also expected an association between the number of matching tweets and the number of fact-checked tweets, as it reasonable to assume that the more tweets belonging to a story get fact-checked, the more tweets in general support the story. As predicted, both (log_{10}-transformed) variables were positively correlated ($r(346) = .223, p = 1.39{\cdot}10^{-5}, \text{one} - \text{tailed}$; see Fig. 1b). However, this correlation might be confounded by the number of subqueries per story, as this measure (partially) depends upon the number of fact-checked tweets (see Methods section). Therefore, we also calculated the semi-partial correlation between both variables while controlling for the effect of the number of subqueries on the number of matching tweets. The association between both variables remained statistically significant ($r(345) = .132, p = 6.75 \cdot 10^{-3}, \text{one} - \text{tailed}$), despite an evident decrease in the correlation coefficient. Again, we only observed a weak correlation between the variables of interest, which might be explained by the lack of a strong correlation between fact-checking efforts and the true sizes of the stories.

Associations between the number of matching tweets and recall

We expected an association between the number of matching tweets and recall because more sensitive queries should also yield more matching tweets. We observed that both variables were positively correlated ($r(346) = .211, p = 3.76{\cdot}10^{-5}, \text{one} - \text{tailed}$; see Fig. 1c), indicating that queries with higher recall also identified more matching tweets. Please note that we log_{10}-transformed the number of matching tweets because of the extreme skewness of the distribution (see Fig. 2f in the main text). Although (highly) significant, the correlation coefficient is considered to be small, which might be explained by the fact that stories arguably differ significantly in their true number of story-supporting tweets. This may add a (not controlled) source of variance to the number of matched tweets, lowering the correlation.

Furthermore, the relationship between recall and the number of matching tweets should also be observable when considering subqueries belonging to a story. As stated above, we identified up to 6 subqueries per story; those subqueries were chosen in a decreasing order in terms of their recall, i.e., the most sensitive subquery was selected first, followed by the second most sensitive subquery and so forth. Therefore, we expected that the number of matching tweets would decrease with increasing subquery position. We used the repeated measures analysis of variance (rmANOVA) to check whether there is a decrease in the number of matching tweets with increasing subquery position. We performed the analysis separately for different numbers of subqueries, i.e., individual models were estimated for stories having at least two, three, four, five, or six subqueries. The results of these models are shown in Table 1. In fact, we observed a decrease in the number of matching tweets for less sensitive subqueries (see Fig. 1d), but this effect diminished when we considered stories with higher numbers of minimal subqueries ($k \geq 4$). The absence of this effect for a higher number of minimal subqueries may be explained by reduced statistical power due to the above-mentioned exponential decrease in the number of stories having at least k subqueries.

Table 1: **rmANOVAs, performed separately for different numbers of subqueries, in the number of matching tweets.**

		rmANOVA					
$k \geq$	N (%)	df_1	df_2	F	p	η_p^2	Significant contrasts
2	215 (61.78)	1	214	4.81	.029	.022	q1 vs. q2 : $t(214) = 2.19, p = .015$
3	98 (28.16)	2	194	3.23	.045	.032	q1 vs. q2 : $t(97) = 2.07, p = .041$ q1 vs. q3 : $t(97) = 2.58, p = .017$
4	44 (12.64)	3	129	1.27	.289	.029	-
5	16 (4.6)	4	60	.251	.869	.016	-
6	4 (1.15)	5	15	1.07	.404	.262	-

Please note that k is the number of subqueries. Also note that η_p^2 refers to the partial eta squared. Sphericity was checked using Mauchly's W; when sphericity was violated, p-values were corrected using the Greenhouse–Geisser correction. For the pairwise tests (one-tailed), we applied the Holm–Bonferroni method to correct for alpha inflation.

Association between the number of matching tweets and the number of baselines

As described in the Methods section, it is reasonable to assume that queries tend to be more precise when using more baselines, thereby lowering the number of tweets matching a query (as the number of false positives decreases). Therefore, we tested the hypothesis that the number of matching tweets for four baselines was reduced relative to two or three baselines. Again, the number of matching tweets was log_{10}-transformed, resulting in a mean for two/three and four baselines of 5.69 and 5.28, respectively. Therefore, we observed an effect of $\beta = .415$, implying a relative change by the factor $10^{\beta} = 10^{.415} = 2.6$, meaning that the expected number of matching tweets for two/three baselines was more than three times as much as for four baselines. Welch's t-test confirmed that the number of matching tweets was reduced for four baselines ($t(307.36) = -3.87, p = 6.56 \cdot 10^{-5}, \mathrm{one-tailed}$; see Fig. 1e). However, this reduction might be driven by confounding effects of decreased recall and/or a lower number of subqueries for stories having four baselines. To account for potential confounding effects, we run a linear regression model (using the STATSMODELS OLS function; see [1]) treating the log_{10}-transformed number of matching tweets as the dependent measure (y), number of baselines as the independent measure (x), and recall and the number of subqueries as covariates (z_1 and z_2). Please note that we pooled stories with two and three baselines into one category because only a very few stories had two baselines. The model was specified by the formula $y \sim x + z1 + z2$. The results of the model are shown in the output 1.

As shown in output 1, the number of matching tweets was significantly reduced for four baselines after controlling for recall and number of subqueries (see coefficient X[T.4 BL]). The linear regression model therefore confirmed that, despite an evident effect reduction, the number of matching tweets was reduced for four baselines, as indicated by the group

Output 1: Results of the linear regression model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            np.log10(y)   R-squared:                       0.231
Model:                            OLS   Adj. R-squared:                  0.224
Method:                 Least Squares   F-statistic:                     34.47
Date:                Fri, 23 Jun 2023   Prob (F-statistic):           1.67e-19
Time:                        08:40:25   Log-Likelihood:                -453.20
No. Observations:                 348   AIC:                             914.4
Df Residuals:                     344   BIC:                             929.8
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.8601      0.114     42.532      0.000       4.635       5.085
x[T.4 bl]     -0.3237      0.100     -3.244      0.001      -0.520      -0.127
z1            -0.0453      0.354     -0.128      0.898      -0.741       0.650
z2             0.3878      0.049      7.965      0.000       0.292       0.484
==============================================================================
Omnibus:                        7.562   Durbin-Watson:                   2.162
Prob(Omnibus):                  0.023   Jarque-Bera (JB):                7.061
Skew:                           0.296   Prob(JB):                       0.0293
Kurtosis:                       2.630   Cond. No.                         19.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

coefficient being significantly different from zero ($\beta = .324$; $t(344) = -3.24, p = 6.46 \cdot 10^{-4}, \mathrm{one} - \mathrm{tailed}$).

References

[1] Seabold S, Perktold J (2010) Statsmodels: Econometric and Statistical Modeling with Python. In: Proceedings of the 9th Python in Science Conference, DOI 10.25080/majora-92bf1922-011